

# An Optimal Method for Covariate Balancing and Its Properties

BY YICHEN QIN, YANG LI, AND FEIFANG HU\*

arXiv:1611.02802v1 [stat.ME] 9 Nov 2016

---

\*Yichen Qin is Assistant Professor, Department of Operations, Business Analytics and Information Systems, University of Cincinnati, Cincinnati, OH 45221 (E-mail: qinyin@ucmail.uc.edu). Yang Li is Associate Professor, School of Statistics and Center for Applied Statistics, Renmin University of China, Beijing, P.R. China, 100872 (E-mail: yang.li@ruc.edu.cn). Feifang Hu is Professor, Department of Statistics, George Washington University, Washington, DC 20052 (E-mail: feifang@gwu.edu). The research was partially supported by NSF Awards (DMS-1442192 and DMS-1612970) and the National Natural Science Foundation of China (No. 11371366 and No. 71301162).

# An Optimal Method for Covariate Balancing and Its Properties

*SUMMARY:* This article introduces a new randomization procedure to improve the covariate balance across treatment groups. Covariate balance is one of the most important concerns for successful comparative studies, such as causal inference and clinical trials, because it reduces bias and improves the accuracy of inference. However, chance imbalance may still exist in traditional randomized experiments, in which units are randomly allocated without using their covariate information. To address this issue, the proposed method allocates the units sequentially and adaptively, using information on the current level of imbalance and the incoming unit's covariate. With a large number of covariates or a large number of units, the proposed method shows substantial advantages over the traditional methods in terms of the covariate balance and computational time, making it an ideal technique in the era of big data. Furthermore, the proposed method attains the optimal covariate balance, in the sense that the estimated average treatment effect under the proposed method attains its minimum variance asymptotically. Numerical studies and real data analysis provide further evidence of the advantages of the proposed method.

*KEYWORDS:* Causal inference, covariate-adaptive randomization, clinical trial, Mahalanobis distance, minimum asymptotical variance, treatment effect.

# 1 INTRODUCTION

Randomization is an essential tool for the accurate evaluation of the treatment effect, because it mitigates selection bias and provides a basis for statistical inference. However, traditional randomization methods, such as complete randomization (CR), often generate unsatisfactory randomization configurations with unbalanced prognostic (or baseline) covariates; this problem has been recognized and extensively discussed ever since [Fisher \(1926\)](#) noted:

*“Most of experimenters on carrying out a random assignment of plots will be shocked to find out how far from equally the plots distribute themselves.”*

Covariate balance is critical for successful statistical inference in comparative studies. Therefore, these chance imbalances in the covariates obtained from traditional randomization methods can significantly undermine the validity of subsequent analysis. According to [Hu et al. \(2014\)](#) and [McEntegart \(2003\)](#), the advantages of balanced covariates are at least threefold. First, covariate balance improves the accuracy and efficiency of statistical inference across treatment groups, removes the bias in estimation, and increases the true power of hypothesis testing. Second, it increases the interpretability of the estimated treatment effect by making the units in the treatment groups more comparable, thereby enhancing the credibility of the analysis. Third, it makes the analysis more robust against model misspecification, because less modeling is needed for balanced treatment groups. In addition, outliers are more likely to cancel each other out, because they tend to be more evenly distributed across treatment groups.

In the absence of covariate balance, various problems must be addressed before a valid conclusion can be drawn. For instance, in causal inference, if a significant imbal-

ance exists, it is very difficult to make a direct comparison across treatment groups. In this case, any inferences regarding the treatment effect will be biased, and any claims about the treatment effect will need to rely on unverifiable assumptions ([Morgan, 2011](#)). In practice, researchers must assess the balance in the covariate distribution even before applying any methods for estimating the causal effect. Although some ex-post adjustments, such as regression ([Freedman, 2008](#)) and subsample selection using matching or trimming based on propensity scores ([Imbens and Rubin, 2015](#)), can be performed to cope with such an imbalance, they are much less efficient than achieving an ex-ante balance from the start ([Bruhn and McKenzie, 2008](#)). In addition, when applying these adjustments, researchers often need to have at least a nearly correct model, which can be difficult to test ([Cochran, 1965](#); [Cochran and Rubin, 1973](#)). [Rubin \(2008\)](#) explained why the greatest possible efforts should be made during the design phase of an experiment rather than during the analysis stage, at which point the researcher has the potential to bias the results (consciously or unconsciously) ([Morgan, 2011](#)). This issue of covariate imbalance has also led to discussions regarding whether randomization is preferable to a purposefully designed balanced allocation ([Gosset, 1938](#); [Greenburg, 1951](#); [Arnold, 1986](#); [Senn, 2004](#)).

Similarly, in clinical trials, if unbalanced covariates are strongly correlated with the outcomes, their presence may make it difficult to interpret the results of statistical tests with regard to the treatment effect. The credibility of the study also comes into question. Although regression adjustments are available for researchers to address the imbalance issue, regression relies on assumptions that are often difficult to verify, such as, linear versus nonlinear models or homoscedasticity versus heteroscedasticity. In some cases, regression estimates are not even unbiased for the treatment effect ([Imbens and Rubin, 2015](#)).

More recently, covariate balance has attracted growing interest in the field of crowdsourced-internet experimentation ([Horton et al., 2011](#); [Chandler and Kapelner, 2013](#); [Kapelner and Krieger, 2014](#)). Researchers increasingly recruit workers from on-line labor markets into their experiments, such as by asking them to label tumor cells in images or classify articles. Because of the nature of the recruiting process, a large number of workers with many covariates (e.g., 2500 workers in [Chandler and Kapelner \(2013\)](#)), typically are enrolled in such studies, which consequently pose challenges for traditional randomization methods. Therefore, ensuring balanced randomization is a pressing task for any statistical inference.

Furthermore, the phenomenon of covariate imbalance is exacerbated as the number of covariates  $p$  and the sample size  $n$  increase, which is nearly ubiquitous in the era of big data. For example, suppose that the probability of one particular covariate being unbalanced is  $\alpha = 5\%$ . For a study with 10 covariates, the chance of at least one covariate exhibiting imbalance is  $1 - (1 - \alpha)^p = 40\%$ . To address the covariate imbalance issue in the framework of causal inference, [Morgan and Rubin \(2012\)](#) have proposed rerandomization (RR), for which the procedure can be summarized as follows:

- (1) Collect covariate data.
- (2) Specify a balance criterion to determine when a randomization is acceptable. For example, the criterion could be defined as a threshold of  $a > 0$  on the Mahalanobis distance between the sample means across different treatment groups,  $M = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T [\text{cov}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^{-1} (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ , where  $\bar{\mathbf{x}}_1$  and  $\bar{\mathbf{x}}_2$  are the sample means for two treatment groups.
- (3) Randomize the units into treatment groups using traditional randomization methods, such as complete randomization (CR).
- (4) Check the balance criterion  $M < a$ . If the criterion is satisfied, go to Step (5);

otherwise, return to Step (3).

- (5) Perform the experiment using the final randomization obtained in Step (4).

They have further demonstrated various desirable properties of the causal inference performed under rerandomization, such as the reduction in variance of the estimated treatment effect. Although rerandomization works well in the case of a few covariates, it is incapable of scaling up to address massive amounts of data. For example, as the number of covariates increases, the probability of acceptance,  $p_a = P(M < a)$ , of each randomization in Step (4) decreases drastically, causing the rerandomization procedure to remain in Steps (3) and (4) for a long time. In addition, as the sample size increases, the computational time required for each iteration of Step (4) grows linearly. Therefore, rerandomization becomes infeasible in these cases.

In this article, we propose a covariate-adaptive randomization (CAM) approach to generate a more balanced treatment allocation and thus to improve the quality of the subsequent causal inference. Unlike rerandomization or complete randomization, in which all units are allocated independently, we allocate units adaptively and sequentially. We first initialize our algorithm by randomly allocating a few units. For the remaining units, we allocate one randomly chosen pair of units at a time. For each pair of units, using their covariate information and the existing level of imbalance of the previously allocated units, we adjust the probability with which the pair is allocated to treatment groups to avoid incidental covariate imbalance. In this way, we are able to produce a much more balanced allocation of units. As demonstrated through the numerical studies, for cases with a large number of covariates or a large number of units, the proposed method exhibits superior performance, with a more balanced randomization and much less computational time. In addition, the proposed method is proven to be optimal, in the sense that the estimated treatment effect under the

proposed method achieves the minimum asymptotic variance.

The proposed method is following the similar spirit of the minimization methods used in clinical trials ([Taves, 1974](#); [Pocock and Simon, 1975](#); [Hu and Hu, 2012](#)). However, the context of these minimization methods is different from ours. In their settings, patients are observed and allocated as they enter into a clinical trial. In our case, we initially possess the complete information about all units, but we choose to allocate them sequentially and adaptively. In addition, in minimization methods, patients are typically allocated individually (one at a time), whereas in the proposed framework, we randomly choose pairs of units to allocate, which is infeasible in the minimization methods' setting. Another significant difference is that the minimization methods are suitable only for discrete covariates, minimizing the margin and stratum imbalance. The proposed method, in contrast, is suitable for allocating units with both discrete and continuous covariates.

This article is organized as follows. We introduce the proposed covariate-adaptive randomization method using the Mahalanobis distance (CAM) in Section 2 and investigate its theoretical properties in Section 3. We demonstrate its advantages in the treatment effect estimation and present its corresponding theoretical properties in Section 4. We further present an example using real data to demonstrate the superior performance of our method in Section 5. Finally, we conclude with a discussion in Section 6 and relegate the outlining of proofs to Section 7.

## 2 COVARIATE-ADAPTIVE RANDOMIZATION VIA MAHALANOBIS DISTANCE

Suppose that  $n$  units (samples) are to be assigned to two treatment groups and that  $n$  is even. Let  $T_i$  be the assignment of the  $i$ -th unit, i.e.,  $T_i = 1$  for treatment 1 and  $T_i = 0$  for treatment 2. Consider  $p$  continuous covariates for each unit. Let  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$  represent the covariates of the  $i$ -th unit, where  $\mathbf{x}_i \in \mathbb{R}^p$ , and  $x_{ij} \in \mathbb{R}$  for  $j = 1, \dots, p$ . The proposed procedure, covariate-adaptive randomization via Mahalanobis distance (CAM), involves the following steps:

- (1) Choose the covariate balance criterion. We define the Mahalanobis distance,  $M \geq 0$ , between the covariate sample means for the two treatment groups as

$$M(n) = np_n(1 - p_n)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T \text{cov}(\mathbf{x})^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

where  $p_n = \sum_{i=1}^n T_i/n$ , and we fix  $p_n = 1/2$  to ensure that we have equal sample sizes across the treatment groups.  $\bar{\mathbf{x}}_1$  and  $\bar{\mathbf{x}}_2$  are the covariate sample means corresponding to these treatment groups under the proposed method, and  $\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2 \in \mathbb{R}^p$ .  $\text{cov}(\mathbf{x}) \in \mathbb{R}^{p \times p}$  is the covariance matrix of the covariate  $\mathbf{x}$ . In practice, the covariance matrix is usually replaced with the sample covariance matrix, which can be computed using all units (before the randomization starts). This Mahalanobis distance functions as a measure of the covariate balance throughout this article. A smaller value of  $M(n)$  indicates a better covariate balance.

Note that in the case of complete randomization (CR), in which all units are independently allocated to treatment groups with equal probabilities, this Mahalanobis distance would become  $M(n) = (\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T [\text{cov}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)]^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$  ([Morgan, 2011](#); [Morgan and Rubin, 2012](#)).



- (2) Randomly arrange all  $n$  units in a sequence  $\mathbf{x}_1, \dots, \mathbf{x}_n$ .
- (3) Separately assign the first two units to treatment 1 and treatment 2.
- (4) Suppose that  $2i$  units have been assigned to treatment groups ( $i \geq 1$ ).
- (5) For the  $(2i + 1)$ -th and  $(2i + 2)$ -th units:
  - (5a) If the  $(2i + 1)$ -th unit is assigned to treatment 1 and the  $(2i + 2)$ -th unit is assigned to treatment 2 (i.e.,  $T_{2i+1} = 1$  and  $T_{2i+2} = 0$ ), then we can calculate the “potential” Mahalanobis distance,  $M_i^{(1)}$ , between the updated treatment groups with  $2i + 2$  units.
  - (5b) Similarly, if the  $(2i + 1)$ -th unit is assigned to treatment 2 and the  $(2i + 2)$ -th unit is assigned to treatment 1 (i.e.,  $T_{2i+1} = 0$  and  $T_{2i+2} = 1$ ), then we can calculate the “potential” Mahalanobis distance,  $M_i^{(2)}$ , between the updated treatment groups with  $2i + 2$  units.
- (6) Assign the  $(2i + 1)$ -th and  $(2i + 2)$ -th units to treatment groups according to the following probabilities:

$$P(T_{2i+1} = 1, T_{2i+2} = 0 | \mathbf{x}_{2i}, \dots, \mathbf{x}_1, T_{2i}, \dots, T_1) = \begin{cases} q & \text{if } M_i^{(1)} < M_i^{(2)}, \\ 1 - q & \text{if } M_i^{(1)} > M_i^{(2)}, \\ 0.5 & \text{if } M_i^{(1)} = M_i^{(2)}, \end{cases}$$

$$P(T_{2i+1} = 0, T_{2i+2} = 1 | \mathbf{x}_{2i}, \dots, \mathbf{x}_1, T_{2i}, \dots, T_1) = 1 - P(T_{2i+1} = 1, T_{2i+2} = 0 | \mathbf{x}_{2i}, \dots, \mathbf{x}_1, T_{2i}, \dots, T_1),$$

$$P(T_{2i+1} = 0, T_{2i+2} = 0 | \mathbf{x}_{2i}, \dots, \mathbf{x}_1, T_{2i}, \dots, T_1) = 0,$$

$$P(T_{2i+1} = 1, T_{2i+2} = 1 | \mathbf{x}_{2i}, \dots, \mathbf{x}_1, T_{2i}, \dots, T_1) = 0,$$

where  $0.5 < q < 1$ .

- (7) Repeat Steps (4) through (6) until all units are assigned.

The value of  $q$  is set to 0.75 throughout this article. Different values of  $q$  will not affect the theoretical results presented in this article. For a further discussion of  $q$ , please see Hu and Hu (2012). In this algorithm,  $n$  is assumed to be even. If  $n$  is odd, then the last ( $n$ -th) unit is randomly assigned to either treatment 1 or 2 with a probability of 0.5.

Note that the units are not necessarily observed sequentially; however, we allocate them sequentially (in pairs) to minimize the occurrence of covariate imbalance. The sequence in which the units are allocated is not unique. Rather, there are  $n!$  different possible sequences, but their performances are similar, especially when  $n$  is large (see Figure 1 for details).

### 3 THEORETICAL PROPERTIES OF COVARIATE-ADAPTIVE RANDOMIZATION VIA MAHALANOBIS DISTANCE

We study the asymptotic properties of the Mahalanobis distance,  $M(n)$ , obtained using the proposed method.  $M(n)$  represents the level of imbalance of the covariates, such that at smaller values of  $M(n)$ , the covariates are more balanced.

**Theorem 3.1.** *Under the proposed method, suppose that the covariate  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , is independent and identically distributed as a multivariate normal distribution with zero mean; then we have  $M(n) = O_p(n^{-1})$ .*

Note that the Mahalanobis distance that is obtained through the complete randomization (CR),  $M_{\text{CR}}(n)$ , has a stationary distribution of a Chi-square distribution

with  $p$  degrees of freedom (regardless of  $n$ ), i.e.,  $M_{\text{CR}}(n) \sim \chi_{df=p}^2$ . Therefore, the Mahalanobis distance that is obtained through the rerandomization (RR),  $M_{\text{RR}}(n)$ , has a Chi-square distribution with  $p$  degrees of freedom conditional on  $M_{\text{RR}}(n) < a$ , i.e.,  $M_{\text{RR}}(n) \sim \chi_{df=p}^2 | \chi_{df=p}^2 < a$ . Hence, as the sample size  $n$  increases, the proposed method reveals a greater advantage over both rerandomization and complete randomization, because  $M(n)$  converges to 0 at the rate of  $1/n$ . That is, under the proposed method, as more units are included in the study, the quality (balance level) of the randomization becomes significantly better.

Moreover, under complete randomization, as the number of covariates  $p$  increases, the stationary distribution of  $M_{\text{CR}}(n)$  becomes flatter, which implies poorer allocation in terms of covariate balance (i.e., larger  $M_{\text{CR}}(n)$ ). As a consequence, rerandomization has a lower probability of acceptance,  $p_a = P(M_{\text{CR}}(n) < a)$ . Therefore, the advantage of the proposed method also becomes more significant as  $p$  increases, because the  $M(n)$  obtained using the proposed method converges to 0 regardless of the magnitude of  $p$ . In addition, even for a fixed  $n$ , the effect of  $p$  on  $M(n)$  under the proposed method is less severe than that under rerandomization or complete randomization (as illustrated later in Figure 1).

To demonstrate the advantage of the proposed method, we have conducted a simulation to compare the proposed method with rerandomization (with a fixed acceptance probability of  $p_a = 0.3$ ) using the multivariate normal covariates with the zero mean and identity covariance matrix; the results are presented in Figure 1. For different sample sizes  $n$  and different numbers of covariates  $p$ , we plot the histograms of  $M(n)$  and  $M_{\text{RR}}(n)$  obtained using the simulated data. As the figure shows, as  $n$  increases with fixed  $p$ , the distribution of the Mahalanobis distance  $M_{\text{RR}}(n)$  obtained through rerandomization remains unchanged, whereas the distribution of the  $M(n)$  obtained using the proposed method rapidly converges to 0. Moreover, as  $p$  increases with fixed  $n$ ,

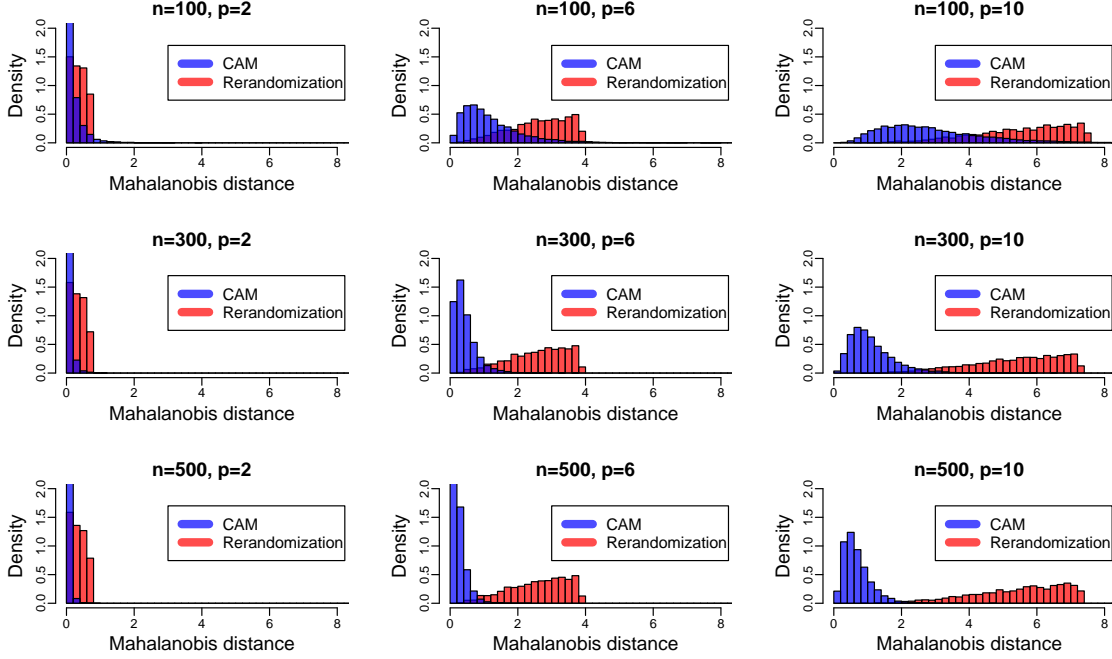


Figure 1: Comparison of the distributions of the Mahalanobis distances obtained via the proposed method,  $M(n)$ , and rerandomization,  $M_{RR}(n)$ , for different sample sizes  $n$  and different numbers of covariates  $p$ .

the distributions obtained through rerandomization and the proposed method become wider, but the inflation of distribution is much less severe for the proposed method (i.e., the overlap between the two distributions becomes smaller as  $p$  increases).

In addition, we also compared the proposed method with rerandomization in terms of computational times and feasibility. Note that the proposed method only requires one iteration (i.e., it processes all units once), whereas rerandomization requires multiple iterations of complete randomization to achieve an acceptable balance level. Therefore, we compared the number of iterations required for rerandomization to achieve the same performance (same Mahalanobis distance) as the proposed method. In addition, we also compared the corresponding computational times. The results are shown in Figure 2. As seen in Figures 2a and 2b, when the sample size and number of covariates

are small, the computational advantage of the proposed method is not obvious. As  $n$  and  $p$  increase, however, the proposed method gradually shows a significant advantage over rerandomization, because more iterations and more time are required for rerandomization in order to achieve the same level of performance as the proposed method. As the number of covariates continues to increase, rerandomization will eventually become infeasible. In other words, it is nearly impossible for rerandomization to achieve the same performance as the proposed method. Note that the computational time of the proposed method grows only linearly with  $n$  and remains the same for different  $p$ s, whereas the computational time of rerandomization grows exponentially as either  $n$  or  $p$  increases.

Finally, we verified the rate of convergence stated in Theorem 3.1 by plotting the expected Mahalanobis distance against the reciprocal of the sample size ( $1/n$ ) using simulated data, as shown in Figure 3. It is clear that for different numbers of covariates  $p$ , the expected Mahalanobis distance converges to 0 at the same rate of  $1/n$ , as evidenced by the straight lines in the figure.

## 4 CAUSAL INFERENCE UNDER COVARIATE-ADAPTIVE RANDOMIZATION AND PROPERTIES

### 4.1 Framework

We use a framework similar to that of [Morgan and Rubin \(2012\)](#). After allocating the units to treatment groups, we are interested in estimating the treatment effect from the outcomes  $y_i$ ,  $i = 1, \dots, n$ , obtained in the treatment groups. Suppose that  $y_i(T_i)$

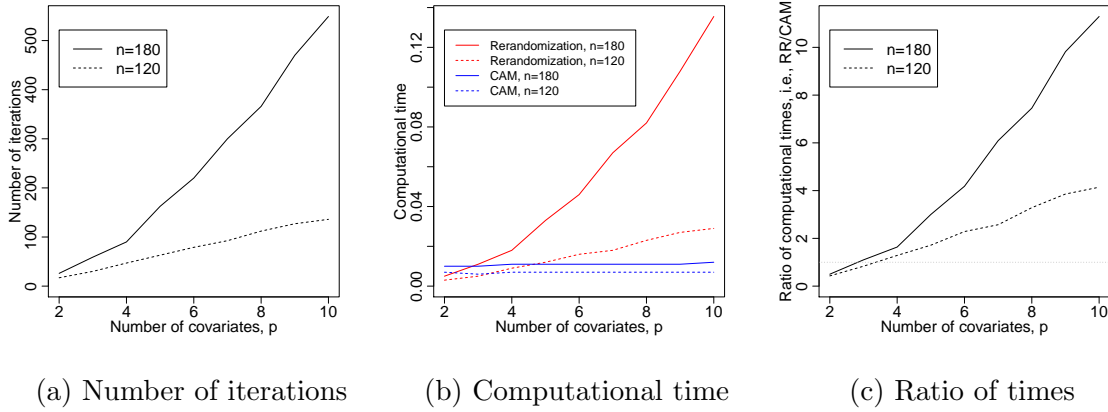


Figure 2: Comparison of the numbers of iterations, the computational times, and the ratios of computational times for rerandomization and the proposed method. Note that the figures show the *medians* (instead of the means) of the numbers of iterations and computational times, because there were many Monte Carlo iterations in which rerandomization required too much time and the simulation had to be terminated. Panel (a): numbers of iterations of rerandomization required to achieve the same performance as the proposed method. Panel (b): the corresponding computational times of the two methods compared in Panel (a). Panel (c): the ratios of computational times shown in Panel (b).

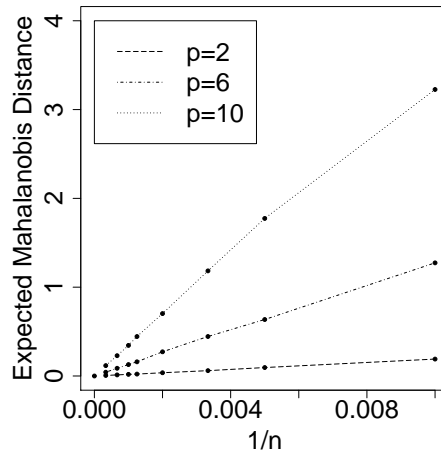


Figure 3: Verification of the rate of convergence of  $M(n)$  using the proposed method.

represents the potential outcome of the  $i$ -th unit under the treatment  $T_i$  (Rubin's causal model, [Rubin \(1974\)](#)). The actual observed outcome  $y_i$  can be expressed as

$$y_i = y_i(1)T_i + y_i(0)(1 - T_i).$$

The average treatment effect is

$$\tau = \frac{\sum_{i=1}^n y_i(1)}{n} - \frac{\sum_{i=1}^n y_i(0)}{n}.$$

However, the fundamental problem in causal inference is that we only observe  $y_i(T_i)$  for one particular  $T_i$ . Therefore,  $\tau$  cannot be calculated directly. If we want to estimate  $\tau$  using  $y_i$ , a natural choice is

$$\hat{\tau} = \frac{\sum_{i=1}^n T_i y_i}{\sum_{i=1}^n T_i} - \frac{\sum_{i=1}^n (1 - T_i) y_i}{\sum_{i=1}^n (1 - T_i)}, \quad (1)$$

which is simply the difference between the sample means of the outcome variable  $y_i$  for the different treatment groups.

One problem with  $\hat{\tau}$  is that if there is an imbalance in the covariates, it will affect the accuracy of  $\hat{\tau}$ . For example, if we estimate the effect of a drug when the treatment 1 group contains mostly males and the treatment 2 group contains mostly females, then the estimated treatment effect  $\hat{\tau}$  will not be able to exclude the effect of gender. In other words, the difference in the covariates will make  $\hat{\tau}$  less accurate.

To adjust for such an imbalance, we can use linear regression to estimate the treatment effect. That is, conditional on the treatment assignment  $T_i$ , each outcome variable is assumed to follow the model below:

$$y_i = \mu_1 T_i + \mu_2 (1 - T_i) + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \epsilon_i, \quad (2)$$

where  $\mu_1$  and  $\mu_2$  are the main effects of treatments 1 and 2, respectively, and  $\mu_1 - \mu_2 = \tau$  is the treatment effect. Furthermore,  $\beta_j$  represents the covariate effect, and  $\epsilon_i$  is an independent and identically distributed random error with zero mean and constant

variance  $\sigma_\epsilon^2$ , and is independent of  $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})^T$ . All covariates  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , are independent and identically distributed.

Let us define

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \mathbf{T} = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{bmatrix}, \tilde{\mathbf{T}} = \begin{bmatrix} T_1 & 1 - T_1 \\ T_2 & 1 - T_2 \\ \vdots & \vdots \\ T_n & 1 - T_n \end{bmatrix},$$

$\widetilde{\mathbf{X}} = [\tilde{\mathbf{T}}; \mathbf{X}]$ ,  $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ , and  $\boldsymbol{\beta}^* = (\mu_1, \mu_2, \beta_1, \dots, \beta_p)^T$ . Then, we can obtain the ordinary least squares estimate of  $\boldsymbol{\beta}^*$ :

$$\hat{\boldsymbol{\beta}}^* = (\widetilde{\mathbf{X}}^T \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T \mathbf{Y}.$$

Let us consider  $\mathbf{L} = (1, -1, 0, \dots, 0)^T$ , a  $(p+2)$ -dimensional vector. We define

$$\tilde{\tau} = \mathbf{L}^T \hat{\boldsymbol{\beta}}^*,$$

which is another estimate of the treatment effect that is adjusted for the imbalance in the covariates. Note that if  $\widetilde{\mathbf{X}}$  does not include any covariates, i.e.,  $\widetilde{\mathbf{X}} = \tilde{\mathbf{T}}$ , then the regression model is  $y_i = \mu_1 T_i + \mu_2 (1 - T_i) + \epsilon_i$ , and in Equation (1),  $\tilde{\tau}$  becomes  $\hat{\tau}$ , which is the estimated treatment effect without adjusting for the imbalance in the covariates.

In the next section, we study the properties of  $\hat{\tau}$  and  $\tilde{\tau}$  under our proposed method (i.e.,  $\hat{\tau}_{\text{CAM}}$  and  $\tilde{\tau}_{\text{CAM}}$ ), under complete randomization (i.e.,  $\hat{\tau}_{\text{CR}}$  and  $\tilde{\tau}_{\text{CR}}$ ), and under rerandomization (i.e.,  $\hat{\tau}_{\text{RR}}$  and  $\tilde{\tau}_{\text{RR}}$ ).

## 4.2 Theoretical Properties of the Estimated Treatment Effect

Morgan and Rubin (2012) proved that under complete randomization and rerandomization,  $\hat{\tau}_{\text{CR}}$  and  $\hat{\tau}_{\text{RR}}$  are unbiased. For the proposed method, we can similarly show unbiasedness:



**Theorem 4.1.** *Under the proposed method, we have*

$$\mathbb{E}[\hat{\tau}_{\text{CAM}}|\mathbf{X}, \text{CAM}] = \tau.$$

In addition to unbiasedness, we can also show the following:

**Theorem 4.2.** *Under the proposed method, suppose that the covariate  $\mathbf{x}_i$ ,  $i = 1, \dots, n$ , is independent and identically distributed as a multivariate normal distribution with zero mean; then we have*

$$\text{cov}[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2|\mathbf{X}, \text{CAM}] = u_n \text{cov}[\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2|\mathbf{X}, \text{CR}],$$

where  $u_n = \mathbb{E}[M(n)|\mathbf{X}, \text{CAM}]$  and  $u_n = O(n^{-1})$ .

In randomized experiments, the emphasis typically is placed on the percent reduction in variance (PRIV), which was originally defined by [Morgan and Rubin \(2012\)](#). This quantity represents the percentage by which the randomization method reduces the variance of the differences in the means calculated for the different treatment groups. Therefore, a higher value of the PRIV indicates that the means are closer to each other. Consider the PRIV for the  $j$ -th covariate,

$$100 \left( \frac{\text{Var}[\bar{x}_{j,1} - \bar{x}_{j,2}|\mathbf{X}, \text{CR}] - \text{Var}[\bar{x}_{j,1} - \bar{x}_{j,2}|\mathbf{X}, \text{CAM}]}{\text{Var}[\bar{x}_{j,1} - \bar{x}_{j,2}|\mathbf{X}, \text{CR}]} \right),$$

where  $\bar{x}_{j,1}$  and  $\bar{x}_{j,2}$  are the  $j$ -th elements of  $\bar{\mathbf{x}}_1$  and  $\bar{\mathbf{x}}_2$ . According to Theorem 4.2, the PRIV of each covariate is  $100(1 - u_n)\%$  under the proposed method. We recall that for rerandomization, the PRIV of each covariate is  $100(1 - v_a)\%$ , which is a constant and independent of the sample size. This is because  $v_a$  is a function of only  $a$ . In contrast, for the proposed method,  $\text{PRIV}_{\text{CAM}} \rightarrow 100\%$  as  $n \rightarrow \infty$ .

Theorem 4.2 implies that the quality of the “matching” is much improved using the proposed method. As the sample size increases, the imbalance in the covariates reaches the minimum level. This is particularly useful when the covariates and outcome are

correlated, because in this case, the proposed method will in turn improve the precision of the estimation of the treatment effect, as detailed in the following theorem.

**Theorem 4.3.** *Under the proposed randomization method, suppose that the outcome variable  $y_i$  and the covariate  $\mathbf{x}_i$  are normally distributed and that the treatment effect is additive; then, the percent reduction in variance (PRIV) of  $\hat{\tau}_{\text{CAM}}$  is  $100(1 - u_n)R^2$ , where  $R^2$  is the squared multiple correlation between  $y_i$  and  $\mathbf{x}_i$  within the treatment groups, and  $u_n = O(n^{-1})$ .*

Notably, Theorem 4.3 does not assume a linear regression for the outcome model.

Let us compare the PRIV of  $\hat{\tau}_{\text{CAM}}$  under the proposed method with the PRIV of  $\hat{\tau}_{\text{RR}}$  under rerandomization. Recall that the PRIV of  $\hat{\tau}_{\text{RR}}$  is  $100(1 - v_a)R^2$  (Morgan and Rubin, 2012), which is again a constant and does not depend on the sample size. In contrast, the PRIV of  $\hat{\tau}_{\text{CAM}}$  is  $100(1 - u_n)R^2$  and converges to  $100R^2$  as the sample size  $n \rightarrow \infty$ . In fact, the PRIV of  $\hat{\tau}_{\text{CAM}}$  is simply the PRIV of the covariates scaled by  $R^2$ . To illustrate the advantage of the proposed method, we plot the PRIVs of  $\hat{\tau}_{\text{CAM}}$  and of  $\hat{\tau}_{\text{RR}}$  (with a fixed acceptance probability of  $p_a = 0.1$ ) in Figure 4. Note that we let  $R^2 = 1$  in both figures only for illustrative purposes (as in Morgan and Rubin (2012)). It is evident that as  $n$  increases, at each value of  $p$ , the PRIV of  $\hat{\tau}_{\text{CAM}}$  increases to 100%. However, the PRIV of  $\hat{\tau}_{\text{RR}}$  at a given  $p$  does not vary with different  $n$ . The advantage of the proposed method over rerandomization is clear, especially for large  $n$  and large  $p$ .

Meanwhile, the percent reduction in variance due to the adjustment via linear regression in complete randomization is  $100[(1 + M_{\text{CR}}(n)/n)R^2 - M_{\text{CR}}(n)/n]$  (Cox, 1982), which converges to  $100R^2$  as  $n \rightarrow \infty$ . Therefore, we conclude that the proposed method can reduce the asymptotic variance to the minimum level.

In addition, if we further assume that the outcome variable  $y_i$  truly follows a linear regression model, we can show that  $\hat{\tau}_{\text{CAM}}$  achieves the optimal precision even without

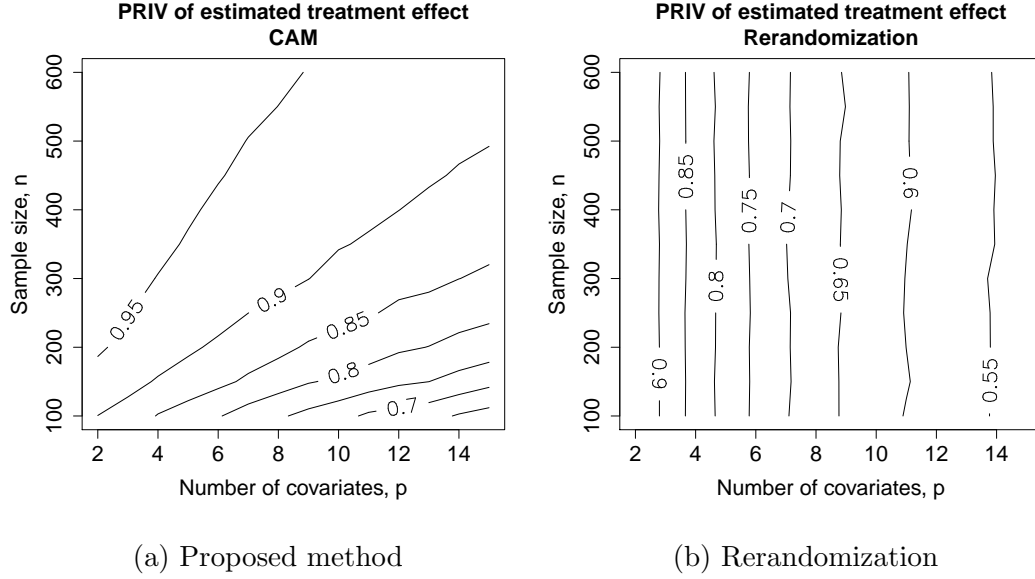


Figure 4: The percent reductions in variance of the estimated treatment effect under the proposed method,  $\hat{\tau}_{\text{CAM}}$ , and under rerandomization,  $\hat{\tau}_{\text{RR}}$ , for various sample sizes and numbers of covariates. Panel (a): proposed method. Panel (b): rerandomization.

adjusting for the imbalance in the covariates using linear regression. That is,

**Theorem 4.4** (Optimal precision). *Suppose that the outcome variable  $y_i$  truly follows the linear regression model in Equation (2) and that we estimate the treatment effect under the proposed method and under complete randomization; then, we have*

$$\begin{aligned}
\sqrt{n}(\hat{\tau}_{\text{CAM}} - (\mu_1 - \mu_2)) &\xrightarrow{D} N(0, V_1), \\
\sqrt{n}(\tilde{\tau}_{\text{CAM}} - (\mu_1 - \mu_2)) &\xrightarrow{D} N(0, V_2), \\
\sqrt{n}(\tilde{\tau}_{\text{CR}} - (\mu_1 - \mu_2)) &\xrightarrow{D} N(0, V_3), \\
\sqrt{n}(\hat{\tau}_{\text{CR}} - (\mu_1 - \mu_2)) &\xrightarrow{D} N(0, V_4),
\end{aligned}$$

where  $4\sigma_\epsilon^2 = V_1 = V_2 = V_3 < V_4$ .

This theorem first implies that under the proposed method, the precision of the estimate of the treatment effect obtained using a simple sample mean difference,  $\hat{\tau}_{\text{CAM}}$ , is

the same as the precision of the estimate obtained through a linear regression which adjusts for the covariate imbalance,  $\tilde{\tau}_{\text{CAM}}$ . This suggests that the regression adjustment would not be necessary under the proposed method, because the covariates already would have been balanced sufficiently well, and the linear regression does not provide any additional benefit. Furthermore, the theorem also implies that the precision of  $\hat{\tau}_{\text{CAM}}$  is the same as the precision of the estimated treatment effect obtained from a linear regression under complete randomization,  $\tilde{\tau}_{\text{CR}}$ , which is considered optimal. In other words, the proposed method can balance the covariates so well that, asymptotically, the simple sample mean difference  $\hat{\tau}_{\text{CAM}}$  is just as good as the linear-regression-adjusted estimate  $\tilde{\tau}_{\text{CR}}$ . Therefore, the adjustment provided by linear regression is generally not needed.

Moreover, under complete randomization, if the outcome variable follows the linear regression model, then the asymptotic variance of the estimated treatment effect,  $\tilde{\tau}_{\text{CR}}$ , reaches its minimum level. Therefore, we conclude that the estimated treatment effect under the proposed method,  $\hat{\tau}_{\text{CAM}}$ , also attains optimal precision, i.e.,  $4\sigma_\epsilon^2$ . We also understand that the Mahalanobis distance that we have selected as the covariate imbalance criterion is the best criterion to use in a randomization procedure for linear regression models. Although  $\tilde{\tau}_{\text{CR}}$  and  $\hat{\tau}_{\text{CAM}}$  have the same precision, it is worth noting that to calculate  $\tilde{\tau}_{\text{CR}}$ , it is necessary to estimate all regression coefficients  $\beta^*$ , whereas  $\hat{\tau}_{\text{CAM}}$  is simply the sample mean difference and does not require the estimation of any additional coefficients. This is especially advantageous when the number of covariates is large, as in the case of high-dimensional data.

In Table 1, we summarize the relationships between the asymptotic variances of the different estimates presented in Theorem 4.4. Note that “Asym. Var.” in the table stands for the asymptotic variances of the different estimates blown up by a factor of  $n$ .

Randomized covariates	Randomization method	Working model for estimating $\mu_1 - \mu_2$	
		$\mathbf{1m}(\mathbf{Y} \sim \tilde{\mathbf{T}})$	$\mathbf{1m}(\mathbf{Y} \sim \tilde{\mathbf{T}} + \mathbf{X})$
$\mathbf{X}$	Complete randomization	Asym. Var. >	Asym. Var.
		$\vee$	$\parallel$
	CAM	Asym. Var. =	Asym. Var.

Table 1: Demonstration of the relationship of asymptotic variances of different estimates in Theorem 4.4.

Randomized covariates	Randomization method	Working model for estimating $\mu_1 - \mu_2$	
		$\mathbf{1m}(\mathbf{Y} \sim \tilde{\mathbf{T}})$	$\mathbf{1m}(\mathbf{Y} \sim \tilde{\mathbf{T}} + \mathbf{X})$
$\mathbf{X}$	Complete randomization	161.1932	144.5853
	CAM	145.5646	145.6051

Table 2: Simulation study for verification of Table 1.

In addition, we performed a simple simulation to verify Table 1, and presented the results in Table 2. In the simulation, we used four continuous covariates in  $\mathbf{X}$  and simulate the outcome according to  $y_i = \mu_1 T_i + \mu_2 (1 - T_i) + 1 * x_{i1} + 1 * x_{i2} + 1 * x_{i3} + 1 * x_{i4} + \epsilon_i$ , where  $\mu_1 = 0$ ,  $\mu_2 = 1$ ,  $x_{ij} \sim N(0, 1)$ , and  $\epsilon_i \sim N(0, 36)$ , and the sample size was  $n = 5000$ .

### 4.3 Computational Time

The previous section clearly demonstrates the advantages of the proposed method with regard to causal inference. A natural question is whether we can also let  $v_a \rightarrow 0$  in the rerandomization method to improve its performance to match that of the

proposed method (because rerandomization allows researchers to increase the power of the analysis at the expense of computational time (Morgan and Rubin, 2012)). However, this option is infeasible in many cases, as illustrated below.

Suppose that the time required to allocate one additional unit using the proposed method is  $Cp$ , where  $p$  is the number of covariates and  $C > 0$ . Suppose that the time required to allocate one additional unit through complete randomization is  $R > 0$ . Then, we have the following theorem.

**Theorem 4.5.** *To achieve the same level of performance, the ratio of the average computational time of the proposed method to the average computational time of the rerandomization method is proportional to  $\chi^2_{df=p}(a^*)Cp/R$ , where  $\chi^2_{df=p}(\cdot)$  is the cumulative distribution function of a Chi-square distribution with  $p$  degrees of freedom, and  $a^*$  is the root of  $\gamma(p/2, a^*/2)Dp = 2\gamma(p/2 + 1, a^*/2)n$  where  $D > 0$  is a constant and  $\gamma(w, t) = \int_0^t x^{w-1} \exp\{-x\}dx$  is the incomplete gamma function.*

To illustrate the advantage of the proposed method, we plot the ratio of the computational times between the proposed method and rerandomization as a function of  $n$  and  $p$  in Figure 5. We let  $C = 10$ ,  $R = 1$ , and  $D = 5$  for illustrative purposes. In Figure 5, we can see that as either  $p$  or  $n$  increases, the ratio of the computational times rapidly approaches 0. When  $n = 500$  and  $p = 15$ , the ratio is  $\exp(-22)$ , which means that it is (almost) impossible for rerandomization to achieve the same level of performance as the proposed method. This demonstration illustrates the advantages of the proposed method in the case of a large sample size or high-dimensional data.

Furthermore, we report the ratios of the computational times for several scenarios as quantitative values in Table 3. We assume  $C = 10$ ,  $R = 1$ , and  $D = 5$ . As we can see, for a small sample size and low-dimensional covariates, the computational times of the proposed method and rerandomization are comparable. However, as either  $p$  or  $n$  increases, the ratio approaches 0 very fast.

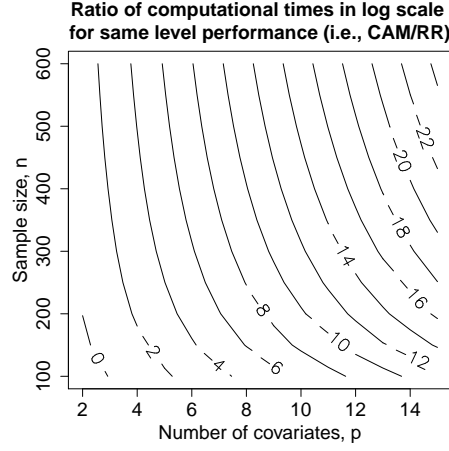


Figure 5: Ratio of the computational time of the proposed method to that of rerandomization (i.e., CAM/RR) for achieving the same level of performance. The values are given on a logarithmic scale.

Sample size $n$	$p = 2$	4	6	8	10	12
200	0.9830	0.1084	0.0094	7.492e-04	5.686e-05	4.197e-06
400	0.4957	0.0275	0.0012	4.884e-05	1.876e-06	7.010e-08
600	0.3312	0.0123	0.0003	9.748e-06	2.510e-07	6.289e-09

Table 3: Ratio of the computational time of the proposed method to that of rerandomization (i.e., CAM/RR) for achieving the same level of performance.

## 5 REAL DATA EXAMPLE

In this section, we illustrate the advantages of our proposed method using a real data set obtained in a clinical study of a Ceragem massage (CGM) thermal therapy bed, a medical device for the treatment of lumbar disc disease. The study was conducted by the medical device company that produces the CGM thermal therapy bed. In total, 186 patients have been chosen for the study. For each patient, there are 50 covariates, i.e.,  $\mathbf{x}_i \in \mathbb{R}^{50}$ . Among the covariates, there are 30 numerical covariates, such as age and several baseline measurements of the patient’s current conditions, including lower back pain, leg pain, leg numbness, body examination scores, and magnitudes of pain in other body parts (shoulders, neck, chest, hip and so on), all measured on 0-10 scales. In the original study, these patients were randomly assigned to the treatment group (to which the thermal therapy was administrated using the medical device) or the control group. At the end of the study, their outcome variable  $y_i$ , representing the numerical measurements of their lower back pain after the treatment, was recorded to study the treatment effect of the CGM thermal therapy bed.

In the original study, because the patients are randomly allocated, the Mahalanobis distance of the patients allocation was only 57.67, which is relatively high and indicates a moderate covariate imbalance across the treatment groups.

Using the same patients’ covariates, we reassigned these patients to treatment and control groups using the proposed method, complete randomization, and rerandomization. We repeated the group allocation for these 186 patients many times, and the Mahalanobis distances obtained using all methods are plotted in Figure 6. We also replicated the data four times to obtain a sample size of  $n = 744 = 186 * 4$ , and then also allocated these 744 patients many times using all the methods. We report the corresponding Mahalanobis distances in the same figure.



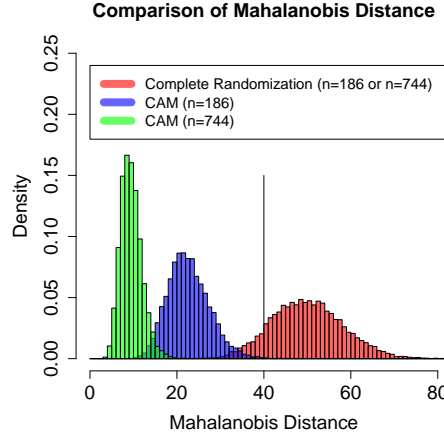


Figure 6: Comparison of the distributions of the Mahalanobis distance obtained using the proposed method, complete randomization, and rerandomization. Note that rerandomization is represented by the portion of the complete randomization distribution that lies to the left of the vertical line ( $M = 40$ ).

As seen from the figure, the Mahalanobis distances obtained by applying the proposed method to the original data ( $n = 186$ ) are consistently lower than those obtained through complete randomization. Hypothetically, if we had  $n = 744$  patients, the Mahalanobis distance could be further decreased toward 0 (i.e., the green histogram). In that case, few of the allocations obtained through complete randomization could achieve the same level of balance as the proposed method. Note that regardless of the sample size ( $n = 186$  or  $n = 744$ ), the Mahalanobis distance obtained through complete randomization always follows the Chi-square distribution. If we set the criterion for rerandomization to  $M < 40$ , then rerandomization will produce the Mahalanobis distances to the left of the vertical line ( $M = 40$ ), which are still not comparable with the results of the proposed method. We could further reduce the threshold to  $M < 30$ ; however, a lower threshold results in a lower acceptance probability. Moreover, please note that the proposed method requires only a single run.

Under each patient allocation scheme, we further simulated the outcome variable

using a linear regression model which was fitted to the original data. To closely mimic the original data, we fitted the linear regression to the original data with the original patient allocation,  $y_i = \mu_1 T_i + \mu_2(1 - T_i) + \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i$ . We stored the residuals,  $\{\hat{\epsilon}_i\}_{i=1}^n$ , and the coefficient estimates,  $\hat{\mu}_1$ ,  $\hat{\mu}_2$ , and  $\hat{\boldsymbol{\beta}}$ , of the fitted linear regression. For the simulated patient allocations  $T_i^{\text{sim}}$ , we simulate the outcome variable,  $y_i^{\text{sim}}$ , according to  $y_i^{\text{sim}} = \hat{\mu}_1 T_i^{\text{sim}} + \hat{\mu}_2(1 - T_i^{\text{sim}}) + \mathbf{x}_i^T \hat{\boldsymbol{\beta}} + \epsilon_i^{\text{sim}}$ , where  $\epsilon_i^{\text{sim}}$  was uniformly sampled from the residuals  $\{\hat{\epsilon}_i\}_{i=1}^n$ .

Using the simulated outcome variable, we estimated the average treatment effect using  $\hat{\tau} = [\sum_{i=1}^n T_i^{\text{sim}} y_i^{\text{sim}}] / \sum_{i=1}^n T_i^{\text{sim}} - [\sum_{i=1}^n (1 - T_i^{\text{sim}}) y_i^{\text{sim}}] / \sum_{i=1}^n (1 - T_i^{\text{sim}})$  under the proposed method, rerandomization, and complete randomization. The resulting performance comparison is summarized in Table 4. As seen from this table, in terms of estimating the treatment effect, the proposed method exhibits the best performance compared with rerandomization and complete randomization. The proposed method yields the largest percent reduction in variance (PRIV) and the lowest variance. For rerandomization, a smaller threshold (i.e.,  $M < 30$  or  $M < 40$ ) results in better performance; however, this comes at the cost of a longer computational time and a lower acceptance probability. Note that the  $R^2$  for the regression fitted to the original data is only 0.33, and because of the finite sample size, the optimal PRIV cannot be achieved. We can see that if we increase the sample size to  $n = 744$ , the PRIV of the proposed method is greatly improved and is very close to optimal, whereas that of the rerandomization method does not improve at all. The gain from the proposed method is quite substantial. This simulation clearly illustrates the potential of the proposed method in the context of big data, in which both the dimension  $p$  and the sample size  $n$  are relatively large.

Sample Size	Method	PRIV	MSE (or Var)	$u_n$ or $v_a$
$n = 186$	CAM	19.7%	0.081	0.502
	Rerandomization ( $M < 30$ )	15.1%	0.085	0.562
	Rerandomization ( $M < 40$ )	12.2%	0.090	0.730
	Complete randomization	-	0.100	-
$n = 744$	CAM	27.4%	0.018	0.205
	Rerandomization ( $M < 30$ )	14.6%	0.021	0.556
	Rerandomization ( $M < 40$ )	10.9%	0.022	0.718
	Complete randomization	-	0.025	-

Table 4: Comparison of the proposed method with rerandomization and complete randomization for real data analysis.

## 6 DISCUSSION

In this article, we have introduced a new randomization procedure for balancing the covariates to improve the accuracy of causal inference. Compared with traditional methods, the proposed method can cope with a large number of covariates, which is especially advantageous in the era of big data, in which high dimensional data is frequently encountered. The proposed method also shows superior performance in terms of computational time. In addition, it achieves optimality under the linear regression framework, in the sense that, asymptotically, the proposed method can balance the covariates so well that the imbalance adjustment provided by linear regression is not needed.

Although the proposed method is different from the minimization methods and other randomization methods used in clinical trials (Wei, 1978; Begg and Iglewicz,

1980; Atkinson, 1982; Smith, 1984a,b), it can be extended to such settings. Instead of selecting a pair of units, we can select only one unit to allocate. However, the behavior of the Mahalanobis distance in such a scenario will be further complicated, because the proportion of the treatment group (i.e.,  $\sum_{i=1}^n T_i/n$ ) then becomes a random variable. We suspect that the allocation procedure should be slightly modified such that both the Mahalanobis distance and the proportion are controlled. In such a scenario, we anticipate that the rate of convergence of the Mahalanobis distance can be further improved. We leave this possibility as a topic for future investigation.

Many other potential directions for further research remain as well. For example, we have shown the optimality of the estimated treatment effect. An extension to hypothesis testing is also of interest (Ma et al., 2015). The optimality of the estimator hints at the most powerful test for the treatment effect. In addition, as the number of covariates increases, it is more efficient to balance only the most important covariates (Morgan and Rubin, 2015); therefore, the selection of the important covariates to balance in our proposed framework is an interesting topic. The proposed method may also be applied to balance important covariates in the field of crowdsourced-internet experimentation.

## 7 APPENDIX

We provide outlines of the key proofs in the Appendix. The supplementary materials contain detailed proofs of all theorems.

*Proof of Theorem 3.1.* We first convert the covariates to canonical form (Rubin and Thomas, 1992). Let  $\Sigma = \text{cov}(\mathbf{x})$  and  $\mathbf{z}_i = \Sigma^{-1/2}\mathbf{x}_i$  where  $\Sigma^{-1/2}$  is the Cholesky square root of  $\Sigma^{-1}$ . Suppose that  $n$  is even. By the assumption of normality,  $\mathbf{z}_i \sim N(0, \mathbf{I})$ ,

and

$$M(n) = np_n(1 - p_n)(\bar{z}_1 - \bar{z}_2)^T(\bar{z}_1 - \bar{z}_2).$$

We further define

$$\mathbf{y}_n = \frac{n}{2}(\bar{z}_1 - \bar{z}_2) = \sum_{i:T_i=1} \mathbf{z}_i - \sum_{i:T_i=0} \mathbf{z}_i,$$

$$\Delta_{n+2} = (-1)^{T_{n+2}}(\mathbf{z}_{n+1} - \mathbf{z}_{n+2}).$$

We can see that  $\{\mathbf{y}_n, \mathbf{y}_{n+2}, \mathbf{y}_{n+4}, \dots\}$  is a Markov process and  $\mathbf{y}_{n+2} = \mathbf{y}_n + \Delta_{n+2}$ . Define the test function  $V(\mathbf{y}_n) = \mathbf{y}_n^T \mathbf{y}_n$ . By denoting  $\mathbb{E}[\cdot | \mathbf{y}_n] = \mathbb{E}_n[\cdot]$ , we have

$$\mathbb{E}_n[V(\mathbf{y}_{n+2})] - V(\mathbf{y}_n) = \mathbb{E}_n[\mathbf{y}_{n+2}^T \mathbf{y}_{n+2}] - \mathbf{y}_n^T \mathbf{y}_n = 2\mathbb{E}_n[\mathbf{y}_n^T \Delta_{n+2}] + \mathbb{E}_n[\Delta_{n+2}^T \Delta_{n+2}],$$

where  $\mathbb{E}_n[\Delta_{n+2}^T \Delta_{n+2}] = \mathbb{E}_n[(-1)^{2T_{n+2}}(\mathbf{z}_{n+1} - \mathbf{z}_{n+2})^T(\mathbf{z}_{n+1} - \mathbf{z}_{n+2})]$  is a positive constant.

For the first term on the right, we have

$$\begin{aligned} \mathbb{E}_n[\mathbf{y}_n^T \Delta_{n+2}] &= \mathbb{E}_n[\mathbf{y}_n^T (-1)^{T_{n+2}}(\mathbf{z}_{n+1} - \mathbf{z}_{n+2})] \\ &= \mathbb{E}_n\left\{\mathbb{E}\left[\mathbf{y}_n^T (-1)^{T_{n+2}}(\mathbf{z}_{n+1} - \mathbf{z}_{n+2}) \middle| \mathbf{z}_{n+1}, \mathbf{z}_{n+2}\right]\right\} \\ &= \mathbb{E}_n\left\{(1 - 2q)|\mathbf{y}_n^T(\mathbf{z}_{n+1} - \mathbf{z}_{n+2})|\right\} \\ &= \mathbb{E}_n\left\{(1 - 2q)|\mathbf{y}_n| |\mathbf{z}_{n+1} - \mathbf{z}_{n+2}| |\cos \theta|\right\} \\ &= (1 - 2q)|\mathbf{y}_n| \mathbb{E}_n[|\mathbf{z}_{n+1} - \mathbf{z}_{n+2}|] \mathbb{E}_n[|\cos \theta|], \end{aligned}$$

where  $\theta$  is the angle between  $\mathbf{y}_n$  and  $\mathbf{z}_{n+1} - \mathbf{z}_{n+2}$ . Note that  $\mathbb{E}_n[|\mathbf{z}_{n+1} - \mathbf{z}_{n+2}|]$  and  $\mathbb{E}_n[|\cos \theta|]$  are two positive constants. Since  $1 - 2q < 0$ , there exist a constant  $b > 0$  and  $c < 0$ , such as when  $|\mathbf{y}_n| > b$ ,  $\mathbb{E}_n[\mathbf{y}_n^T \Delta_{n+2}] + \mathbb{E}_n[\Delta_{n+2}^T \Delta_{n+2}] < c$ . Therefore,  $\mathbb{E}_n[V(\mathbf{y}_{n+2})] - V(\mathbf{y}_n) < c$  for  $|\mathbf{y}_n| > b$ . Similarly, we have  $\mathbb{E}_n[V(\mathbf{y}_{n+2})] - V(\mathbf{y}_n) < \mathbb{E}_n[\Delta_{n+2}^T \Delta_{n+2}]$  for  $|\mathbf{y}_n| \leq b$ . By the “drift conditions” (Meyn and Tweedie, 2009), we know  $\mathbf{y}_n$  has a stationary distribution. Therefore,  $nM/(4p_n(1 - p_n)) = \mathbf{y}_n^T \mathbf{y}_n$  has a stationary distribution and  $M = O_p(n^{-1})$ .  $\square$

*Proof of Theorem 4.1.* Since  $T_i$  and  $1 - T_i$  are exchangeable, for any  $i$ ,  $\mathbb{E}[T_i|\mathbf{x}_n] = \mathbb{E}[1 - T_i|\mathbf{x}_n]$ . Since  $\mathbb{E}[T_i|\mathbf{x}_n] + \mathbb{E}[1 - T_i|\mathbf{x}_n] = 1$ , we have  $\mathbb{E}[T_i|\mathbf{x}_n] = \mathbb{E}[1 - T_i|\mathbf{x}_n] = 0.5$ .

Suppose  $n$  is even.

$$\begin{aligned}
\mathbb{E}[\hat{\tau}|\mathbf{x}_1, \dots, \mathbf{x}_n] &= \mathbb{E}\left[\frac{\sum_{i=1}^n T_i y_i}{\sum_{i=1}^n T_i} - \frac{\sum_{i=1}^n (1 - T_i) y_i}{\sum_{i=1}^n (1 - T_i)} \middle| \mathbf{x}_1, \dots, \mathbf{x}_n\right] \\
&= \mathbb{E}\left[\frac{\sum_{i=1}^n T_i y_i(1)}{\sum_{i=1}^n T_i} - \frac{\sum_{i=1}^n (1 - T_i) y_i(0)}{\sum_{i=1}^n (1 - T_i)} \middle| \mathbf{x}_1, \dots, \mathbf{x}_n\right] \\
&= \frac{\sum_{i=1}^n \mathbb{E}[T_i|\mathbf{x}_1, \dots, \mathbf{x}_n] y_i(1)}{n/2} - \frac{\sum_{i=1}^n (1 - \mathbb{E}[T_i|\mathbf{x}_1, \dots, \mathbf{x}_n]) y_i(0)}{n/2} \\
&= \frac{\sum_{i=1}^n 1/2 y_i(1)}{n/2} - \frac{\sum_{i=1}^n (1 - 1/2) y_i(0)}{n/2} \\
&= \tau.
\end{aligned}$$

□

*Proof of Theorem 4.2.* Please see supplementary materials.

□

*Proof of Theorem 4.3.* Please see supplementary materials.

□

*Proof of Theorem 4.4.* We first convert the covariates to canonical form (Rubin and Thomas, 1992). Let  $\Sigma = \text{cov}(\mathbf{x})$  and  $\mathbf{z}_i = \Sigma^{-1/2} \mathbf{x}_i$  where  $\Sigma^{-1/2}$  is the Cholesky square root of  $\Sigma^{-1}$ . Suppose  $n$  is even. By the assumption of normality,  $\mathbf{z}_i \sim N(0, \mathbf{I})$ .

Define

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \vdots \\ \mathbf{z}_n^T \end{bmatrix}, \mathbf{T} = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_n \end{bmatrix}, \tilde{\mathbf{T}} = \begin{bmatrix} T_1 & 1 - T_1 \\ T_2 & 1 - T_2 \\ \vdots & \vdots \\ T_n & 1 - T_n \end{bmatrix},$$

and  $\tilde{\mathbf{Z}} = [\tilde{\mathbf{T}}; \mathbf{Z}]$ ,  $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_p)^T = (\Sigma^{-1/2})^T \boldsymbol{\beta}$ ,  $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$  and  $\boldsymbol{\gamma}^* = (\boldsymbol{\mu}^T, \boldsymbol{\gamma}^T)^T = (\mu_1, \mu_2, \gamma_1, \dots, \gamma_p)^T$ .

Then true model, equation (2), can be rewritten as

$$\mathbf{Y} = \tilde{\mathbf{X}} \boldsymbol{\beta}^* + \boldsymbol{\epsilon} = \tilde{\mathbf{T}} \boldsymbol{\mu} + \mathbf{X} \boldsymbol{\beta} + \boldsymbol{\epsilon} = \tilde{\mathbf{T}} \boldsymbol{\mu} + \mathbf{Z} \boldsymbol{\gamma} + \boldsymbol{\epsilon} = \tilde{\mathbf{Z}} \boldsymbol{\gamma}^* + \boldsymbol{\epsilon}.$$

**Part I:**  $\hat{\tau}_{\text{CAM}}$

Suppose  $\mathbf{K} = (1, -1)$ , then  $\hat{\tau}_{\text{CAM}}$  can be obtained by running the regression,  $\mathbf{Y} = \tilde{\mathbf{T}}\boldsymbol{\mu} + \boldsymbol{\epsilon}$ , even though the true model is  $\mathbf{Y} = \tilde{\mathbf{Z}}\boldsymbol{\gamma}^* + \boldsymbol{\epsilon} = \tilde{\mathbf{T}}\boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$ . In particular, we can write  $\hat{\tau}_{\text{CAM}}$  as

$$\begin{aligned}
\hat{\tau}_{\text{CAM}} &= \frac{\sum_{i=1}^n T_i y_i}{\sum_{i=1}^n T_i} - \frac{\sum_{i=1}^n (1 - T_i) y_i}{\sum_{i=1}^n (1 - T_i)} \\
&= \mathbf{K} \left( \frac{\tilde{\mathbf{T}}^T \tilde{\mathbf{T}}}{n} \right)^{-1} \frac{\tilde{\mathbf{T}}^T \mathbf{Y}}{n} \\
&= \mathbf{K} \left( \frac{\tilde{\mathbf{T}}^T \tilde{\mathbf{T}}}{n} \right)^{-1} \frac{\tilde{\mathbf{T}}^T (\tilde{\mathbf{Z}}\boldsymbol{\gamma}^* + \boldsymbol{\epsilon})}{n} \\
&= \mathbf{K} \left( \frac{\tilde{\mathbf{T}}^T \tilde{\mathbf{T}}}{n} \right)^{-1} \frac{\tilde{\mathbf{T}}^T (\tilde{\mathbf{T}}\boldsymbol{\mu} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon})}{n} \\
&= \mathbf{K} \left[ \boldsymbol{\mu} + \left( \frac{\tilde{\mathbf{T}}^T \tilde{\mathbf{T}}}{n} \right)^{-1} \frac{\tilde{\mathbf{T}}^T (\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon})}{n} \right] \\
&= \mu_1 - \mu_2 + \mathbf{K} \left( \frac{\tilde{\mathbf{T}}^T \tilde{\mathbf{T}}}{n} \right)^{-1} \frac{\tilde{\mathbf{T}}^T (\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon})}{n}.
\end{aligned}$$

We know, as  $n \rightarrow \infty$ ,

$$\frac{\tilde{\mathbf{T}}^T \tilde{\mathbf{T}}}{n} \xrightarrow{p} \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix} = \mathbf{M}.$$

We further define

$$\begin{aligned}
\mathbf{A} &= \mathbf{K} \mathbf{M}^{-1} \left[ \frac{\tilde{\mathbf{T}}^T (\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon})}{n} \right], \\
\mathbf{B} &= \mathbf{K} \left[ \left( \frac{\tilde{\mathbf{T}}^T \tilde{\mathbf{T}}}{n} \right)^{-1} - \mathbf{M}^{-1} \right] \left[ \frac{\tilde{\mathbf{T}}^T (\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon})}{n} \right],
\end{aligned}$$

so that  $\hat{\tau}_{\text{CAM}} = \mathbf{A} + \mathbf{B}$ .

For  $\mathbf{A}$ , with some algebra, we can show

$$\mathbf{A} = \frac{2}{n} \left[ \sum_{j=1}^p \sum_{i=1}^n (2T_i - 1) \gamma_j z_{i,j} + \sum_{i=1}^n (2T_i - 1) \epsilon_i \right].$$

For the first term on the right, we have

$$\sum_{j=1}^p \sum_{i=1}^n (2T_i - 1) \gamma_j z_{i,j} = \sum_{j=1}^p \gamma_j \left[ \sum_{i \in \{i: T_i=1\}} z_{i,j} - \sum_{i \in \{i: T_i=0\}} z_{i,j} \right].$$

where  $\{i : T_i = 1\}$  and  $\{i : T_i = 0\}$  represent the two treatment groups. From the proof of Theorem 3.1, we understand that  $\sum_{i \in \{i: T_i=1\}} z_{i,j} - \sum_{i \in \{i: T_i=0\}} z_{i,j}$  is a stationary process under the proposed method (i.e. a mean reverting process as  $n \rightarrow \infty$ ). Therefore,

$$\begin{aligned} \sum_{i \in \{i: T_i=1\}} z_{i,j} - \sum_{i \in \{i: T_i=0\}} z_{i,j} &= O_p(1), \\ \sum_{j=1}^p \gamma_j \left[ \sum_{i \in \{i: T_i=1\}} z_{i,j} - \sum_{i \in \{i: T_i=0\}} z_{i,j} \right] &= O_p(1). \end{aligned}$$

In addition, note that  $(2T_i - 1)^2 = 1$ , we have

$$\begin{aligned} \text{Var} \left( \frac{2}{n} \sum_{i=1}^n (2T_i - 1) \epsilon_i \right) &= \mathbb{E} \left( \frac{4}{n^2} \sum_{i=1}^n (2T_i - 1)^2 \epsilon_i^2 \right) \\ &= \mathbb{E} \left( \frac{4}{n^2} \sum_{i=1}^n \epsilon_i^2 \right) \\ &= \frac{4\sigma_\epsilon^2}{n}. \end{aligned}$$

Therefore,

$$\sqrt{n} \mathbf{A} \xrightarrow{D} N(0, 4\sigma_\epsilon^2).$$

Similarly, for  $\mathbf{B}$ , we will show  $\sqrt{n} \mathbf{B} \xrightarrow{p} 0$ . First note that

$$\left( \frac{\tilde{\mathbf{T}}^T \tilde{\mathbf{T}}}{n} \right)^{-1} - \mathbf{M}^{-1} \xrightarrow{p} 0.$$

Therefore, showing  $\sqrt{n} \mathbf{B} \xrightarrow{p} 0$  is equivalent to show

$$\frac{\tilde{\mathbf{T}}^T (\mathbf{Z} \boldsymbol{\gamma} + \boldsymbol{\epsilon})}{\sqrt{n}} = O_p(1).$$



First, notice that

$$\frac{\tilde{\mathbf{T}}^T(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon})}{\sqrt{n}} = \frac{1}{\sqrt{n}} \begin{bmatrix} \sum_{j=1}^p \sum_{i=1}^n T_i \gamma_j z_{i,j} + \sum_{i=1}^n T_i \epsilon_i \\ \sum_{j=1}^p \sum_{i=1}^n (1 - T_i) \gamma_j z_{i,j} + \sum_{i=1}^n (1 - T_i) \epsilon_i \end{bmatrix}.$$

Since

$$\begin{aligned} \frac{1}{\sqrt{n}} \left( \sum_{j=1}^p \sum_{i=1}^n T_i \gamma_j z_{i,j} + \sum_{i=1}^n T_i \epsilon_i \right) &= \frac{1}{2} \left[ \frac{1}{\sqrt{n}} \left( \sum_{j=1}^p \sum_{i=1}^n \gamma_j z_{i,j} + \sum_{i=1}^n \epsilon_i \right) + \right. \\ &\quad \left. \frac{1}{\sqrt{n}} \left( \sum_{j=1}^p \sum_{i=1}^n (2T_i - 1) \gamma_j z_{i,j} + \sum_{i=1}^n (2T_i - 1) \epsilon_i \right) \right]. \end{aligned}$$

By central limit theorem, we have

$$\frac{1}{\sqrt{n}} \left( \sum_{j=1}^p \sum_{i=1}^n \gamma_j z_{i,j} + \sum_{i=1}^n \epsilon_i \right) = O_p(1).$$

In addition,

$$\frac{1}{\sqrt{n}} \left( \sum_{j=1}^p \sum_{i=1}^n (2T_i - 1) \gamma_j z_{i,j} + \sum_{i=1}^n (2T_i - 1) \epsilon_i \right) = \frac{\sqrt{n}\mathbf{A}}{2}.$$

Since  $\sqrt{n}\mathbf{A}$  converges to a normal distribution,

$$\frac{1}{\sqrt{n}} \left( \sum_{j=1}^p \sum_{i=1}^n (2T_i - 1) \gamma_j z_{i,j} + \sum_{i=1}^n (2T_i - 1) \epsilon_i \right) = O_p(1).$$

Therefore,

$$\frac{1}{\sqrt{n}} \left( \sum_{j=1}^p \sum_{i=1}^n T_i \gamma_j z_{i,j} + \sum_{i=1}^n T_i \epsilon_i \right) = O_p(1).$$

By symmetry, we have

$$\frac{1}{\sqrt{n}} \left( \sum_{j=1}^p \sum_{i=1}^n (1 - T_i) \gamma_j z_{i,j} + \sum_{i=1}^n (1 - T_i) \epsilon_i \right) = O_p(1).$$

Therefore,

$$\frac{\tilde{\mathbf{T}}^T(\mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon})}{\sqrt{n}} = O_p(1).$$

Hence,  $\sqrt{n}\mathbf{B} \xrightarrow{p} 0$ , together with  $\sqrt{n}\mathbf{A} \xrightarrow{D} N(0, 4\sigma_\epsilon^2)$ , by Slutsky's theorem, we have

$$\sqrt{n}(\hat{\tau}_{\text{CAM}} - (\mu_1 - \mu_2)) \xrightarrow{D} N(0, 4\sigma_\epsilon^2).$$

For  $\hat{\tau}_{\text{CR}}$ ,  $\tilde{\tau}_{\text{CAM}}$ , and  $\tilde{\tau}_{\text{CR}}$ , we can obtain their asymptotic distributions in similar ways. Please see supplementary materials for details.  $\square$

*Proof of Theorem 4.5.* Please see supplementary materials.  $\square$

## REFERENCES

- Arnold, G. C. (1986). Randomization: A historic controversy. *The Fascination of statistics*, 1(1):231–244.
- Atkinson, A. C. (1982). Optimum biased coin designs for sequential clinical trials with prognostic factors. *Biometrika*, 69:61–67.
- Begg, C. B. and Iglewicz, B. (1980). A treatment allocation procedure for sequential clinical trials. *Biometrics*, 36(1):81–90.
- Bruhn, M. and McKenzie, D. (2008). In pursuit of balance: Randomization in practice in development field experiments. *World Bank Policy Research Working Papers*.
- Chandler, D. and Kapelner, A. (2013). Breaking monotony with meaning: Motivation in crowdsourcing markets. *Journal of Economic Behavior & Organization*, 90:123–133.
- Cochran, W. G. (1965). The planning of observational studies of human populations. *Journal of the Royal Statistical Society, Series A*, 128(2):234–266.
- Cochran, W. G. and Rubin, D. B. (1973). Controlling bias in observational studies: a review. *Sankhya, A*, 35(4):417–446.
- Cox, D. R. (1982). Randomization and concomitant variables in the design of experiments. *Statistics and Probability: Essays in Honor of C. R. Rao*, pages 197–202. North-Holland, Amsterdam. MR0659470.
- Fisher, R. A. (1926). The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513.
- Freedman, D. A. (2008). On regression adjustments to experimental data. *Advances in Applied Mathematics*, 40(2):180–193.

- Gosset, W. J. (1938). Comparison between balanced and random arrangements of field plots. *Biometrika*, 29(3-4):363–379.
- Greenburg, B. G. (1951). Why randomize? *Biometrics*, 7(4):309–322.
- Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, 14(3):399–425.
- Hu, F., Hu, Y., Ma, Z., and Rosenberger, W. F. (2014). Adaptive randomization for balancing over covariates. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(4):288–303.
- Hu, Y. and Hu, F. (2012). Asymptotic properties of covariate-adaptive randomization. *Annals of Statistics*, 40(3):1794–1815.
- Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.
- Kapelner, A. and Krieger, A. (2014). Matching on-the-fly: Sequential allocation with higher power and efficiency. *Biometrics*, 70(2):378–388.
- Ma, W., Hu, F., and Zhang, L. (2015). Testing hypotheses of covariate-adaptive randomized clinical trials. *Journal of the American Statistical Association*, 110(510):669–680.
- McEntegart, D. J. (2003). The pursuit of balance using stratified and dynamic randomization techniques: An overview. *Therapeutic Innovation & Regulatory Science*, 37(3):293–308.
- Meyn, S. P. and Tweedie, R. L. (2009). *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition edition.
- Morgan, K. L. (2011). Rerandomization to improve covariate balance in randomized experiments. *Ph.D. Thesis, Harvard University*.
- Morgan, K. L. and Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *Annals of Statistics*, 40(2):1263–1282.
- Morgan, K. L. and Rubin, D. B. (2015). Rerandomization to balance tiers of covariates. *Journal of the American Statistical Association*, 110(512):1412–1421.
- Pocock, S. J. and Simon, R. (1975). Sequential treatment assignment with balancing for prognostic factors in the controlled clinical trial. *Biometrics*, 31(1):103–115.
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(688).
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Annals of Applied Statistics*, 2(3):808–840.

- Rubin, D. B. and Thomas, N. (1992). Affinely invariant matching methods with ellipsoidal distributions. *Annals of Statistics*, 20(2):1079–1093.
- Senn, S. (2004). Controversies concerning randomization and additivity in clinical trials. *Statistics in Medicine*, 23(24):3729–3753.
- Smith, R. L. (1984a). Properties of biased coin designs in sequential clinical trials. *Annals of Statistics*, 12(3):1018–1034.
- Smith, R. L. (1984b). Sequential treatment allocation using biased coin designs. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):519–543.
- Taves, D. R. (1974). Minimization: A new method of assigning patients to treatment and control groups. *Clinical Pharmacology & Therapeutics*, 15(5):443–453.
- Wei, L. J. (1978). An application of an urn model to the design of sequential controlled clinical trials. *Journal of the American Statistical Association*, 73:559–563.